Extracting Social Determinants of Health with Large Language Models: A Survey of Clinical NLP Methods, Ethics, and Deployment

Anonymous ACL submission

Abstract

Despite accounting for almost half of health outcome variance, social determinants of health (SDOH), encompassing socioeconomic, environmental, and behavioral factors, remain challenging to extract from clinical text. We present the first comprehensive survey of LLM-driven SDOH extraction, examining how large language models can address this critical extraction challenge while introducing new ethical considerations. Synthesizing over 80 peer-reviewed studies to chart the field's evolution from rule-based systems to modern generative models, our analysis reveals that transformer-based approaches consistently outperform earlier machine learning methods, with parameter-efficient techniques like prompt tuning and retrieval-augmented generation making these advances feasible under clinical resource constraints. However, we identify critical gaps: most research lacks essential bias audits, privacy protections, and hallucination controls required for clinical deployment. While emerging ethical frameworks show promise, their adoption remains limited. We consolidate best practices for reproducible SDOH extraction and highlight key challenges, including multilingual coverage, cross-institutional generalization, and cost-effective deployment. This survey provides both a technical road-map and an ethical framework for advancing SDOH extraction toward safe, responsible clinical integration.

1 Introduction

1.1 The SDOH Extraction Challenge

Social determinants of health (SDOH), including socioeconomic, environmental, and behavioral conditions, account for 30–55% of morbidity and mortality variance and up to 80–90% of modifiable risk in high-income countries (Bhavnani et al., 2023; Magnan, 2017). These factors, education, housing, employment, substance use, and neighborhood environment, influence life expectancy and cause tens of thousands of preventable deaths annually in the United States (Magnan, 2017). Since most SDOH signals appear in EHR free-text (Guevara et al., 2024; Hatef et al., 2021), scalable and accurate extraction is essential for public health, reimbursement, and clinical support. However, manual abstraction is costly and error-prone, underscoring the need for automated methods that are both technically robust and ethically responsible. To address these extraction challenges, large language models (LLMs) have transformed clinical NLP by offering unprecedented capabilities for understanding nuanced medical language and complex reasoning. Yet LLMs introduce critical new challenges around hallucination, bias amplification, and deployment complexity that are particularly dangerous for SDOH extraction, where errors can perpetuate health disparities. Automation often replicates the same limitations of bias and inaccuracy, highlighting the need for careful design to solve, rather than merely shift, manual approaches' challenges.

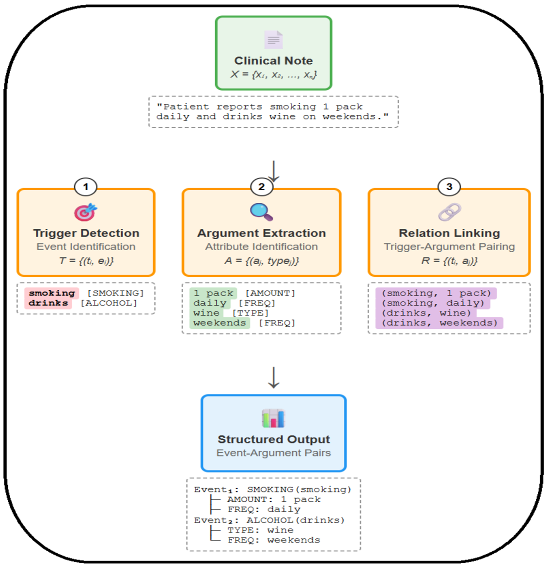

Figure 1: SDOH Event-Based Task Formulation

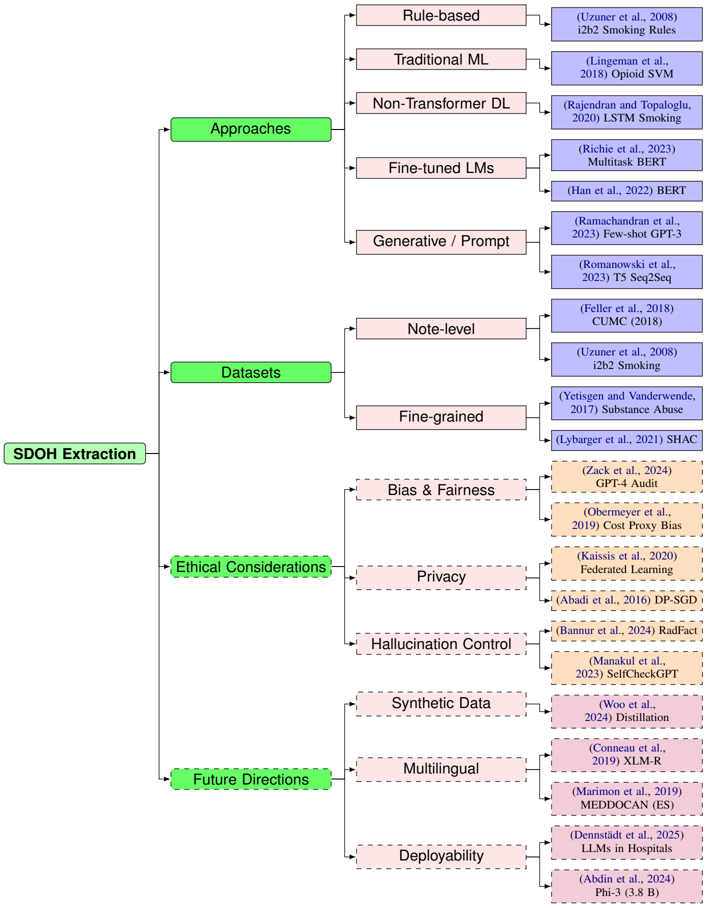

Figure 2: Survey-centric taxonomy: methodological approaches, dataset types, key ethical considerations, and forward-looking research directions for SDOH extraction.

1.2 Study Selection

We conducted a PRISMA-inspired review (Hutton et al., 2016), querying **PubMed, ACL Anthology, IEEE Xplore, ACM Digital Library,** and **Google Scholar**, which yielded 2,595 articles (See Appendix A). After removing duplicates, non-research content, and inaccessible papers, 438 remained for title and abstract screening. The final set included 81 peer-reviewed studies covering rule-based to modern LLMs for SDOH extraction, including works on ethics, bias, and hallucination in clinical NLP. The literature search, conducted between February and June 2025 (covering publications up to May 15, 2025), employed Boolean queries such as `“social determinants of health”` `AND` (`“NLP”` `OR` `“information extraction”`), `“SDOH”` `AND` (`“transformer”` `OR` `“GPT”` `OR` `“BERT”`), `“clinical extraction”` `AND` (`“bias”` `OR` `“privacy”` `OR` `“hallucination”`), etc. We included studies using ML or LLMs for extracting SDOH from clinical free-text with empirical results, excluding those limited to structured data or without methodological contributions. While many papers discuss ethics and bias in SDOH, few propose concrete solutions. To address this gap, we also reviewed broader `“clinical extraction”` research for transferable methodologies.

1.3 Related Work

Early surveys on SDOH extraction focused on rule-based and classical machine learning methods. Patra et al., Bompelli et al., and Li et al. reviewed foundational NLP techniques and broader extraction efforts, while Rajwal et al. proposed protocols for organizing fragmented literature. Disease-specific reviews in orthopedics (Lans et al., 2023), cardiovascular disease (Zhao et al., 2021; McNeill et al., 2023), sickle cell disease (Khan et al., 2023), and mental health (Scherbakov et al., 2025) highlighted limited attention to SDOH, often reduced to demographic proxies. However, these works largely predate LLMs and overlook key issues such as hallucination control, privacy, fairness, and deployment across multilingual or cross-institutional settings.

Task Type	Method (Papers)	Dataset (Performance)	Score	Limitations
Note-level Classification	Rule-based (Uzuner et al., 2008)	i2b2 smoking (502 notes)	0.80–0.89 micro-F1	Misses implicit mentions ("quit ten years ago"); requires exact keywords
	SVM with TF-IDF (Lingeman et al., 2018)	UMass opioid misuse notes	0.81 accuracy	Poor cross-institutional transfer; features don't generalize
	LSTM (Rajendran and Topaloglu, 2020)	6,298 progress notes (smoking)	0.80 F1 (+8% vs SVM)	Long document degradation; struggles with compiled notes
	Flan-T5 XL + UMLS (Gong et al., 2025)	MIMIC-SDoH (5,328 notes)	0.88 macro-F1	Rare categories (<50 examples) still underperform
	Few-shot GPT-3.5 (Consoli et al., 2024)	MIMIC-SDBH, Suicide/Sleep notes	0.90+ AUROC	High lexical diversity categories need more annotations
	LLM pipeline (Gu et al., 2025)	Mass General Brigham EHRs	0.60+ macro-F1	Fails on implicit reasoning; invalid outputs need post-processing
Sentence-level	Flan-T5 XXL (Guevara et al., 2024)	Radiotherapy corpus (6 SDoH)	0.71 macro-F1	No cross-sentence context; imbalanced classes
Sequence Labeling	Bi-LSTM-CRF (Lybarger et al., 2021)	SHAC (4,480 sections)	0.82–0.93 micro-F1	Cannot capture multi-sentence SDoH mentions
	RoBERTa NER (Lituiev et al., 2023)	CLBP corpus (626 notes)	0.84 F1	Drops on rare categories; misses lexical variants
Event-based Extraction	mSpERT (Lybarger et al., 2023a)	SHAC/n2c2 (events)	0.86 F1 overall	Heterogeneous status descriptions cause confusion
	T5-Large seq2seq (Romanowski et al., 2023)	SHAC (12 categories)	0.90 F1	Ambiguous outputs; needs constrained decoding
	GatorTron-GPT (Peng et al., 2023a, 2024)	n2c2/UW challenge	0.84 F1	20B params infeasible; black-box for auditing
	One-shot GPT-4 (Ramachandran et al., 2023)	SHAC test set	0.652 F1	High prompt sensitivity; non-conforming outputs

Table 1: Evolution of SDOH extraction methods: task-specific performance and limitations on benchmark datasets

Problem Definition. Although SDOH are key drivers of health outcomes, most remain buried in unstructured clinical notes. Manual extraction is costly and unreliable, while automated methods face technical challenges (e.g., ambiguous language, institutional variation) and ethical concerns (e.g., bias, privacy). This gap hinders large-scale analysis and reinforces health disparities, calling for solutions that are accurate, scalable, and ethically sound.

Contributions and Road-map. To address this bottleneck, this survey: (i) synthesizes SDOH extraction research through an LLM-centric lens, mapping model types, prompting strategies, and generative paradigms (§2.5, §2.6, §2.7); (ii) compares their efficiency, scalability, and cross-institutional robustness across datasets (§2.2, §4.3); (iii) identifies ethical priorities like bias, fairness, privacy, and hallucination control while formalizing best practices for reproducibility and FAIR data sharing (§3, §4); and (iv) outlines open challenges in multilingual support, data scarcity, and cost-aware deployment (§5). By integrating technical and ethical insights, we offer a roadmap for building responsible, deployable SDOH extraction systems.

2 Foundational Approaches

Figure 2 summarizes the taxonomy of approaches, datasets, ethics, and future directions. Solid boxes indicate SDOH-specific work; dashed boxes show broader ethical and emerging areas. Implementation details appear in Appendix C with a concise overview in Table 1.

2.1 Task Formulation and Evaluation

Research on SDOH extraction from clinical text has evolved through several task formulations. Early work framed it as *note/sentence level classification*, assigning binary (presence/absence of a determinant) or multi-label tags to documents or sentences based on the presence of social factors (Afshar et al., 2019; Jonnagaddala et al., 2015; Feller et al., 2018; Stemerman et al., 2021; Han et al., 2022). Later studies shifted to more granular *information extraction* approaches. In *sequence labeling*, tokens are tagged using the BIO schema to identify SDOH concepts and attributes (Yu et al., 2024). *Relation classification* then links extracted attributes (e.g., frequency, duration) to core concepts. More recent *event-based* methods identify trigger expressions and extract structured arguments (e.g., status, temporality) to represent full SDOH events as illustrated in Figure 1 (Lybarger et al., 2023a; Romanowski et al., 2023). These methods enable richer analyses, including temporal reasoning and patient-level summarization. SDOH extraction systems are typically evaluated using standard information-extraction metrics: micro-averaged precision, recall, and F_1, which reflect the effectiveness of these fine-grained approaches.

2.2 Annotated Corpora for Benchmarking

Many datasets have supported progress in SDOH modeling, though access constraints and skewed label distributions persist. Early corpora focused on smoking status or phenotype mentions, while later datasets added span-level SDOH annotations across various domains like social work, chronic pain, and COVID-19 case reports (Uzuner et al., 2008; Gehrmann et al., 2018; Feller et al., 2020; Wang et al., 2015; Lybarger et al., 2023b, 2021; Han et al., 2022; Lituiev et al., 2023; Raza et al., 2023). However, the under-representation of factors like housing and legal needs, along with institutional access restrictions, hinders generalizability and collaboration. A summary of various SDOH-related datasets is presented in the Appendix B.

2.3 Rule-based and Classical ML Baselines

Initial SDOH systems relied on **rule-based pipelines** using lexical cues, section headers, and negation triggers, achieving strong performance for simple detection tasks but struggling with more complex attribute extraction (Uzuner et al., 2008; Hatef et al., 2019; Bettencourt-Silva et al., 2020; Patra et al., 2021; Bejan et al., 2017; Green et al., 2019; Mowery et al., 2017). Later, **feature-based classifiers** such as linear SVMs and random forests were used with TF-IDF, UMLS concepts, and sentiment features, improving portability while reducing manual rules (Topaz et al., 2019; Wang et al., 2015; Perron et al., 2019; Badger et al., 2019; Amrit et al., 2017; Erickson et al., 2018; Feller et al., 2018). However, these models often overfit and perform poorly under domain shifts, motivating the shift to deep sequence models.

2.4 Recurrent and Deep-learning Encoders

RNN families. The shift to deep learning replaced manual features with automatic representation learning. Recurrent neural networks (RNNs), particularly LSTMs, enabled token-level modeling and captured short-range temporal cues missed by rule-based systems. An LSTM classifier outperformed a TF–IDF SVM by 8 F_1 points for smoking-status detection on progress-note snippets (Rajendran and Topaloglu, 2020), but standard LSTMs struggled with long documents and label dependencies.

Bi-LSTM+CRF with pre-trained embeddings. Bi-LSTM+CRF models initialized with domain-specific embeddings showed substantial gains. On the **SHAC** corpus, such models achieved micro-F_1 scores of **0.82–0.93** across 12 SDOH categories (Lybarger et al., 2021), outperforming prior SVM and rule-based approaches. Incorporating BIOCLINICALBERT embeddings (Alsentzer et al., 2019) further strengthened the baseline in the 2022 n2c2/UW shared task, reducing the gap to transformer-based systems.

Despite these gains, RNNs are inherently sequential, limiting GPU parallelism and degrading on long spans typical of discharge summaries (Song et al., 2018). Their fixed context further hinders cross-sentence reasoning, which is critical for capturing SDOH mentions spread across multiple sentences or paragraphs. These limitations motivated a transition to transformer-based models.

2.5 Transformers and Fine-tuned LLMs

Domain-adaptive BERT variants. Transformer models pre-trained on biomedical corpora transformed SDOH extraction. BIOCLINICALBERT, trained on MIMIC-III and PubMed, improved performance on the SHAC corpus by modeling domain-specific semantics (Alsentzer et al., 2019). Richie et al. (2023) showed that this model outperformed Bi-LSTM+CRF in 12 of 15 SDOH categories (Lybarger et al., 2021), despite using a simpler architecture. BIOBERT, trained on PubMed and PMC, reached $F_1 = 0.92$ on a custom dataset and generalized well across tasks (Lee et al., 2020; Raza et al., 2023).

Scaling up to clinical LLMs. Training on billion-token corpora further improved extraction. GATORTRON-MRC and GATORTRONGPT-20B achieved F_1 scores of 0.74 and 0.84 respectively on the n2c2/UW dataset using decoder-only approaches (Peng et al., 2023b, 2024). T5-style models, such as constrained-decoding T5-large, reached $F_1 = 0.90$ on SHAC (Romanowski et al., 2023). These results reflect transformers' ability to model long-range context while leveraging massive domain-specific pretraining.

Transformer encoders surpass RNNs in scalability and contextual comprehension, but their complexity and resource demands raise challenges in deployment, interpretability, and privacy.

2.6 In-context Prompting and PEFT

Zero and few-shot prompting. Instruction-tuned LLMs demonstrate strong in-context learning, enabling information extraction with only natural-language instructions and a few examples (Brown et al., 2020). On the SHAC test set, a one-shot GPT-4 prompt achieved $F_1 = 0.652$, matching the 7^{th}-ranked supervised system in the n2c2 shared task while requiring no access to private training data (Ramachandran et al., 2023). Broader benchmarks confirm that ChatGPT (OpenAI, 2022), Flan-T5 (Chung et al., 2024), UL2 (Tay et al., 2022), Tk-Instruct (Wang et al., 2022), and Alpaca (Taori et al., 2023) already approach fine-tuned baselines in zero/few-shot settings (Labrak et al., 2023).

Soft prompting and LoRA. Parameter-efficient fine-tuning (PEFT) offers a practical middle ground for institutional deployment. Instead of updating the full model, methods like soft prompting and LoRA adapt only a small set of parameters. Peng et al. (2024) applied soft prompting to a frozen 20B-parameter GATORTRONGPT, achieving strong generalization across institutions and disease cohorts with minimal tuning overhead.

Clinical implications. Prompt-only workflows keep protected health information within the firewall, while PEFT enables lightweight, on-prem deployment under tight resource budgets. Combined with RAG (§2.7), these techniques make LLM-based SDOH extraction feasible for institutions with limited compute capacity. Nevertheless, they raise risks around hallucination and bias amplification, further discussed in §3.

2.7 RAG and Resource Efficiency

Retrieval-augmented generation (RAG) (Lewis et al., 2020) (elaborated in Appendix D) reduces hallucinations (see §3.3) and token costs by limiting model input to relevant snippets from a clinical corpus. Chunk-based retrieval enables GPT-4o (Hurst et al., 2024), Llama-2 (Touvron et al., 2023), and Mistral (Jiang et al., 2023) to match full-note performance on surgical-complication classification with over 50% token savings (Cheetirala et al., 2025; Jiang, 2024). Entity-guided RAG, as in the CLEAR pipeline, improves F_1 to 0.90 (vs. 0.79–0.86 for baselines) while reducing input size and latency by 5× (Lopez et al., 2025). RAG also benefits SDOH tasks. A GPT-4 pipeline yields up to 0.99 precision and 0.88 recall for substance use mentions (Shah-Mohammadi and Finkelstein, 2024). Small models can compete when combined with RAG, e.g., a 2B GEMMA with LoRA matches a 13B Llama-2 on social-note classification when both use CLEAR (Team et al., 2024; Lopez et al., 2025). For zero-label settings, synthetic QA generation with Llama-3 (70B) enables an 8B student to reach micro-$F_1 \geq 0.94$ on clinical tasks (Woo et al., 2024; Grattafiori et al., 2024), while GPT-turbo can produce SDOH-style notes for evaluation of rare categories like unstable housing (Gong et al., 2025). Together, these methods support accurate, efficient, and privacy-conscious extraction pipelines deployable even under hardware and data-sharing constraints.

In summary, (1) RAG reduces cost and hallucination by limiting context, (2) retrieval combined with PEFT allows small local models to match large cloud LLMs, and (3) synthetic distillation fills data gaps when HIPAA (U.S. Department of Health & Human Services, 2003) restricts annotation. These methods together enable scalable, ethical SDOH extraction even in low-resource clinical settings.

3 Ethical Considerations

While LLMs have significantly improved SDOH extraction, these technical advances also introduce critical risks that threaten fair and trustworthy deployment. In this section, we discuss three core concerns: bias, privacy, and hallucination.

3.1 Bias and Fairness

Ensuring fairness across populations is essential for clinical NLP systems. Algorithmic bias is well documented: for example, Obermeyer et al. (2019) found a model that underestimated Black patients' needs by using healthcare spending as a proxy. LLMs may inherit and amplify such disparities due to skewed training data (Zack et al., 2024). Recent tools like LANGFAIR quantify output-side harms, showing that larger models (7B–70B) generate up to 2-4× more harmful outputs for minoritized groups (Bouchard, 2024). While few studies explicitly assess bias in SDOH extraction, mitigation strategies from broader clinical AI, such as balanced sampling, adversarial training, and subgroup F_1 reporting, are applicable. We recommend that future work report per-demographic scores and release audit scripts to support bias assessment.

3.2 Privacy-Preserving Techniques

Because clinical notes are *protected health information*, direct data sharing is often prohibited. This has spurred research into privacy-preserving methods for SDOH modeling. Three main approaches are emerging. **Differential privacy** adds noise during training to provide formal guarantees (Abadi et al., 2016), with recent work showing $F_1 \approx 0.82$ under ($\varepsilon = 0.5$) on clinical text (Henderson and Pearson, 2025). **Federated learning** keeps data local while sharing model weights across institutions (Kaissis et al., 2020), but its application to SDOH remains nascent. **Synthetic data** from generative models like Llama-3.1 can yield strong downstream performance without real data, achieving up to 0.94 micro-F_1 on clinical eligibility tasks (Woo et al., 2024).

3.3 Hallucination Risk and Factuality Control

The same generative strengths that make LLMs effective for SDOH extraction also pose a key risk: the tendency to produce plausible but incorrect outputs, or *hallucinations*, which are dangerous in clinical settings. GPT-3.5 and GPT-4 hallucinated 39.6% and 28.6% of citations, respectively, in a systematic review generation task (Chelli et al., 2024). In a review of 12,999 LLM-generated clinical note sentences, 1.47% (191) were hallucinated, with 44% deemed major and potentially harmful (Asgari et al., 2024). Fabrications were the most common (43%), often found in Plan, Assessment, and Symptoms sections. Although these studies are not specific to SDOH, the risks apply. Systems must guard against factual errors using strategies outlined in Appendix E.1. For SDOH extraction, we recommend: (i) grounding outputs in source spans, (ii) validating with factuality scorers (e.g., RadFact, SelfCheckGPT), and (iii) including a lightweight reviewer interface for clinical verification.

To ensure that such safeguards are reliable across settings, we now turn to the broader challenges of reproducibility and generalization in SDOH extraction.

4 Reproducibility and Generalization

Robust SDOH extraction hinges on reproducibility and generalization, yet most studies overlook them. Drawing on established clinical NLP practices, we outline key strategies here and offer a concise guideline in Appendix F.

4.1 Code and Data Availability

Reliable SDOH extraction requires technical rigor, ethical safeguards, and transparent, reproducible workflows. Since July 2023, ACL's ARR mandates a reproducibility checklist including code, seeds, hyperparameters, and environment details.[1] Yet clinical NLP often lags behind. An audit of seven frameworks found only two met over 50% of 40 criteria; most lacked version control, documentation, or preprocessing metadata (Digan et al., 2021). Similarly, Magnusson et al. (2023) found only 46% of ACL/EMNLP/NAACL 2020–21 papers truly open-sourced their code.

Best practices include: (i) open-sourcing full pipelines via Docker or Conda, (ii) documenting data provenance and licenses in `README` files, (iii) capturing environment details and seed values, and (iv) hosting on GitHub and archival sites like Zenodo. These practices support replicability and, for SDOH, help validate bias and privacy methods before real-world use.

4.2 Dataset Standardization and the FAIR Principles

As models advance, data infrastructure must keep pace. EHR heterogeneity hinders cross-site SDOH modeling, since hospitals store social-history notes in incompatible schemas. Two efforts address this issue:

Common Data Models. The **Observational Medical Outcomes Partnership** Common Data Model (OMOP CDM) standardizes health data into uniform tables and vocabularies (Overhage et al., 2012). Mapping EHRs to OMOP enables shared analytics and OHDSI toolchain use. Zhou et al. (2025) used sentence-transformer embeddings to map free-text medications to OMOP concepts, outperforming string matching. Similar work is now aligning SDOH mentions (e.g., "Lives_with_Mother", "Homeless_Shelter") to SNOMED-CT.[2].

[1] See `https://aclrollingreview.org/faq`
[2] `https://www.snomed.org/`

FAIR Data Stewardship. FAIR principles, Findable, Accessible, Interoperable, Reusable, complement OMOP (Wilkinson et al., 2016). FAIR SDOH corpora should include: *persistent IDs* (e.g., DOIs), *rich metadata* (e.g., source, note type, schema), *standard ontologies* (e.g., SNOMED-CT, LOINC[3]), and clear *data-usage & licensing terms*.

Combining OMOP-like schemas with FAIR practices improves transfer learning, supports new SDOH categories, and enables robust cross-institutional collaboration.

4.3 Cross-Institutional Generalization and Domain Adaptation

SDOH systems must generalize across institutions to ensure equitable performance, but models trained on one site often fail elsewhere due to documentation, terminology, and population differences. This fragility risks amplifying healthcare disparities. Domain-adaptive pretraining addresses this issue. The DRAGON benchmark, covering 28 tasks across five Dutch hospitals, shows consistent gains when using clinical-domain pretraining over general-domain models (Bosma et al., 2025). Effective methods include **continued pretraining (CPT)**, **invariant representation learning**, and **lightweight meta-learning**, detailed in Appendix F. A practical pipeline involves CPT on local unlabeled notes, LoRA or PEFT fine-tuning on a few hundred labeled examples, and validation on a held-out site to detect domain shift.

4.4 Standardized Evaluation Protocols

Reliable SDOH extraction needs a standardized evaluation. Earlier clinical NLP relied on private splits and ad-hoc metrics, hindering fair comparison. Three trends address this:

Shared-task benchmarks. From i2b2 2008 to the 2022 n2c2/UW SDOH Shared Task, organizers now provide task definitions, fixed train/test splits, and official scorers using micro-precision, recall, and F_1 (Lybarger et al., 2023b), discouraging metric cherry-picking.

Multi-site stress tests. DRAGON uses blind evaluation on sequestered data from multiple centers. Though average scores are high, performance varied: 18 of 28 tasks rated excellent/good, 6 moderate, 4 poor, highlighting domain-shift sensitivity (Bosma et al., 2025).

Reporting guidance. ACL's reproducibility checklist advises publishing macro and micro F_1 with 95% CIs via bootstrap. The 2023 n2c2 report complies, sharing its bootstrap script (Lybarger et al., 2023b).

Future SDOH work should use public splits, report macro/micro metrics with CIs, and release scoring code for exact replication. These practices build trust and support fair, ethical innovation.

5 Open Challenges and Future Directions

Building on the technical, ethical, and deployment challenges outlined above, we highlight open problems and future directions to advance SDOH extraction. These reflect persistent gaps in current methods and opportunities for more robust, equitable, and scalable solutions.

5.1 Leveraging LLMs for Data Augmentation and Complex SDOH Representations

Synthetic augmentation and parameter-efficient adaptation. Annotated social-history corpora remain small, limited further by privacy constraints. LLMs offer a way forward by generating synthetic data without exposing protected health information (PHI) as discussed in §2.7. For SDOH, Peng et al. (2024) used prompt-tuning with GatorTronGPT for effective cross-institution and cross-disease transfer, achieving F_1 ∼0.84–0.87 with only soft prompts. These results, however, need validation to confirm linguistic and clinical fidelity.

MRC and seq2seq for richer outputs. SDOH extraction needs richer outputs than flat spans, like capturing attributes, temporal qualifiers, and cross-sentence links (e.g., *lost housing → duration: six months*). Machine-reading comprehension (MRC) based methods let models answer targeted questions per facet (presence, status, duration, etc.), producing structured outputs within clinical LLMs (Peng et al., 2024). These methods show promise but require validation across workflows for clinical utility.

Next steps. For future SDOH extraction systems, research should pursue:

(i) Quality-controlled synthetic pipelines: Use retrieval-conditioned prompts and fact-checkers (e.g., SelfCheckGPT) to ensure clinical authenticity and reduce hallucinations and bias.

(ii) Unified event–argument schema: Extend SHAC with nested attributes and timelines; train seq2seq or MRC models to generate fully linked SDOH graphs interpretable by clinicians.

[3] https://loinc.org/

These directions tackle data scarcity and oversimplified outputs. But technical advances must be paired with fairness, privacy, and utility evaluations to ensure trustworthy, actionable deployment.

5.2 Expanding to Multilingual SDOH Extraction

Despite progress in English SDOH extraction, global deployment demands addressing the field's English-only bias. Most shared tasks and benchmarks exclude the majority of EHRs written in other languages, risking greater inequities. Three core challenges stand out, as elaborated in Appendix G: lack of non-English annotated corpora (e.g., MEDDOCAN (Marimon et al., 2019), DRAGON (Bosma et al., 2025)), cultural and institutional mismatches in SDOH terminology, and limited domain-specific LLMs for languages beyond English. Even state-of-the-art English models struggle when applied cross-lingually.

Recent work offers promising directions. **Translate-train, original-test** approaches achieve comparable F_1 scores of $\sim$0.78-0.79 with careful design of translation pipelines (Fontaine et al., 2023). **Domain-specific pretraining,** such as MedRoBERTa.nl on Dutch notes (Verkijk and Vossen, 2021; Muizelaar et al., 2024), outperforms translated English models on substance use categories. **Synthetic data** from French clinical LLMs (Hiebel et al., 2023) also yields competitive NER performance. These findings underscore the need for culturally aligned resources and multilingual models. A more detailed analysis of the implementation of each of these techniques is provided in Appendix G.1.

Progress depends on community benchmarks and FAIR corpora. In the absence of large resources, multilingual SDOH research must rely on small, vetted datasets and integrate data generation, language adaptation, and cross-national collaboration.

5.3 Towards Responsible and Deployable Clinical LLM Solutions

Technical progress in SDOH extraction must translate into systems that are accurate, trustworthy, and feasible in real-world settings.

Compute vs capability. Many high-performing models, like GPT-4o and Med-PaLM 2, are too large for clinical deployment ($\geq$ 50B parameters). Compact models such as Phi-3 (3.8B) (Abdin et al., 2024) and Gemma-2B (Team et al., 2024), when instruction-tuned (e.g., via LoRA), achieve near-competitive ($\sim$2-3 points) F_1 scores within resource limits. A LoRA-tuned 2B Gemma even matched a 13B Llama-2 on social-work notes (Peng et al., 2024).

On-prem & hybrid deployment. Hospitals favor on-prem or hybrid deployment to safeguard privacy. Cloud APIs offer speed but raise concerns over vendor dependence and data exposure (Dennstädt et al., 2025). Hybrid systems are local retrieval with small-model inference that balance performance, cost, and privacy for SDOH use cases.

Governance and Oversight. Reliable SDOH deployment demands strong governance to ensure fairness and accountability. Tools like LangFair and SelfCheckGPT (§3.3) are being integrated into CI/CD workflows to flag bias and factuality issues pre-deployment. Given the equity implications of SDOH systems, the 2024 PrivateNLP workshop (Habernal et al., 2024) proposed a three-tier model: *risk assessment*, *technical guardrails*, and ongoing *human audit*, aligning with FDA software as a medical device (SaMD) guidance and pointing to potential regulatory pathways.

6 Conclusion

This survey reveals SDOH extraction at a critical juncture where technical progress must align with ethical imperatives for real-world impact. Our review shows that while transformer-based models have advanced extraction capabilities, a fundamental gap remains between research innovation and clinical deployment. Three insights emerge: (i) parameter-efficient LLMs enable advanced methods under clinical constraints; (ii) ethical safeguards are core design requirements, not optional; and (iii) the field's English-centric, institution-specific focus limits global health equity. Progress requires alignment across technical, ethical, and practical fronts, including multilingual benchmarks, built-in fairness controls, and reproducible frameworks for cross-institutional deployment. Ultimately, the value of SDOH extraction will be measured not by F_1 scores alone, but by its capacity to reduce disparities and improve clinical decision-making. The goal is to equip clinicians to act on patients' social context with the same rigor as lab results, demanding a shift from prototypes to responsible, equity-centered systems that transform how healthcare addresses social drivers of health.

7 Limitations

This survey does not include every existing study on SDOH extraction, as some works may fall outside the scope of our focus or lack sufficient methodological detail for meaningful comparison. Instead, we present a curated set of representative papers that align with our research questions, covering key task formulations, evaluation practices, and model innovations. While the survey is not exhaustive, it captures the central developments and challenges in the field, offering a coherent and focused overview to guide future research.

Grammar checking and LaTeX formatting were assisted by automated tools. The authors are solely responsible for the final content and analysis.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Majid Afshar, Andrew Phillips, Niranjan Karnik, Jeanne Mueller, Daniel To, Richard Gonzalez, Ron Price, Richard Cooper, Cara Joyce, and Dmitriy Dligach. 2019. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *Journal of the American Medical Informatics Association*, 26(3):254–261.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Chintan Amrit, Tim Paauw, Robin Aly, and Miha Lavric. 2017. Identifying child abuse through text mining and machine learning. *Expert systems with applications*, 88:402–418.

Vladimir Araujo, Andres Carvallo, Carlos Aspillaga, and Denis Parra. 2020. On adversarial examples for biomedical nlp tasks. *arXiv preprint arXiv:2004.11157*.

Elham Asgari, Nina Montana-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. 2024. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv*, pages 2024–09.

Jonathan Badger, Eric LaRose, John Mayer, Fereshteh Bashiri, David Page, and Peggy Peissig. 2019. Machine learning for phenotyping opioid overdose events. *Journal of biomedical informatics*, 94:103185.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Cosmin A Bejan, John Angiolillo, Douglas Conway, Robertson Nash, Jana K Shirey-Rice, Loren Lipworth, Robert M Cronin, Jill Pulley, Sunil Kripalani, Shari Barkin, and 1 others. 2017. Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *Journal of the American Medical Informatics Association: JAMIA*, 25(1):61.

Joao H Bettencourt-Silva, Natalia Mulligan, Marco Sbodio, John Segrave-Daly, Richard Williams, Vanessa Lopez, and Carlos Alzate. 2020. Discovering new social determinants of health concepts from unstructured data: framework and evaluation. In *Digital Personalized Health and Medicine*, pages 173–177. IOS Press.

Suresh K Bhavnani, Weibin Zhang, Daniel Bao, Mukaila Raji, Veronica Ajewole, Rodney Hunter, Yong-Fang Kuo, Susanne Schmidt, Monique R Pappadis, Elise Smith, and 1 others. 2023. Subtyping social determinants of health in all of us: network analysis and visualization approach. *Medrxiv*.

Anusha Bompelli, Yanshan Wang, Ruyuan Wan, Esha Singh, Yuqi Zhou, Lin Xu, David Oniani, Bhavani Singh Agnikula Kshatriya, Joyce Joy E Balls-Berry, and Rui Zhang. 2021. Social and behavioral determinants of health in the era of artificial intelligence with electronic health records: A scoping review.

Joeran S Bosma, Koen Dercksen, Luc Builtjes, Romain André, Christian Roest, Stefan J Fransen, Constant R Noordman, Mar Navarro-Padilla, Judith Lefkes, Natália Alves, and 1 others. 2025. The dragon benchmark for clinical nlp. *npj Digital Medicine*, 8(1):1–10.

Dylan Bouchard. 2024. An actionable framework for assessing bias and fairness in large language model use cases. *arXiv preprint arXiv:2407.10853*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Satya Narayana Cheetirala, Ganesh Raut, Dhavalkumar Patel, Fabio Sanatana, Robert Freeman, Matthew A Levin, Girish N Nadkarni, Omar Dawkins, Reba

Miller, Randolph M Steinhagen, and 1 others. 2025. Less context, same performance: A rag framework for resource-efficient llm-based clinical nlp. *arXiv preprint arXiv:2505.20320*.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: comparative analysis. *Journal of medical Internet research*, 26:e53164.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Bernardo Consoli, Xizhi Wu, Song Wang, Xinyu Zhao, Yanshan Wang, Justin Rousseau, Tom Hartvigsen, Li Shen, Huanmei Wu, Yifan Peng, and 1 others. 2024. Sdoh-gpt: Using large language models to extract social determinants of health (sdoh). *arXiv preprint arXiv:2407.17126*.

Fabio Dennstädt, Janna Hastings, Paul Martin Putora, Max Schmerder, and Nikola Cihoric. 2025. Implementing large language models in healthcare while balancing control, collaboration, costs and security. *NPJ digital medicine*, 8(1):143.

William Digan, Aurélie Névéol, Antoine Neuraz, Maxime Wack, David Baudoin, Anita Burgun, and Bastien Rance. 2021. Can reproducibility be improved in clinical natural language processing? a study of 7 clinical nlp suites. *Journal of the American Medical Informatics Association*, 28(3):504–515.

Jennifer Erickson, Kenneth Abbott, and Lucinda Susienka. 2018. Automatic address validation and health record review to identify homeless social security disability applicants. *Journal of Biomedical Informatics*, 82:41–46.

Daniel J Feller, Oliver J Bear Don't Walk, Jason Zucker, Michael T Yin, Peter Gordon, Noémie Elhadad, and 1 others. 2020. Detecting social and behavioral determinants of health with structured and free-text clinical data. *Applied clinical informatics*, 11(01):172–181.

Daniel J Feller, Jason Zucker, Bharat Srikishan, Roxana Martinez, Henry Evans, Michael T Yin, Peter Gordon, Noémie Elhadad, and 1 others. 2018. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 422.

Xavier Fontaine, Félix Gaschi, Parisa Rastin, and Yannick Toussaint. 2023. Multilingual clinical ner: Translation or cross-lingual transfer? *arXiv preprint arXiv:2306.04384*.

Sebastian Gehrmann, Franck Dernoncourt, Yeran Li, Eric T Carlson, Joy T Wu, Jonathan Welt, John Foote Jr, Edward T Moseley, David W Grant, Patrick D Tyler, and 1 others. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS one*, 13(2):e0192360.

Aidan Gilson, Xuguang Ai, Thilaka Arunachalam, Ziyou Chen, Ki Xiong Cheong, Amisha Dave, Cameron Duic, Mercy Kibe, Annette Kaminaka, Minali Prasad, and 1 others. 2024. Enhancing large language models with domain-specific retrieval augment generation: A case study on long-form consumer health question answering in ophthalmology. *arXiv preprint arXiv:2409.13902*.

Lei Gong, Jaren Bresnick, Aidong Zhang, Cathy Wu, and Kishlay Jha. 2025. Boosting social determinants of health extraction with semantic knowledge augmented large language model. In *AMIA Annual Symposium Proceedings*, volume 2024, page 453.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Carla A Green, Nancy A Perrin, Brian Hazlehurst, Shannon L Janoff, Angela DeVeaugh-Geiss, David S Carrell, Carlos G Grijalva, Caihua Liang, Cheryl L Enger, and Paul M Coplan. 2019. Identifying and classifying opioid-related overdoses: A validation study. *Pharmacoepidemiology and Drug Safety*, 28(8):1127.

Bowen Gu, Vivian Shao, Ziqian Liao, Valentina Carducci, Santiago Romero Brufau, Jie Yang, and Rishi J Desai. 2025. Scalable information extraction from free text electronic health records using large language models. *BMC Medical Research Methodology*, 25(1):23.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, and 1 others. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ivan Habernal, Sepideh Ghanavati, Abhilasha Ravichander, Vijayanta Jain, Patricia Thaine, Timour Igamberdiev, Niloofar Mireshghallah, and Oluwaseyi

Feyisetan, editors. 2024. *Proceedings of the Fifth Workshop on Privacy in Natural Language Processing*. Association for Computational Linguistics, Bangkok, Thailand.

Sifei Han, Robert F Zhang, Lingyun Shi, Russell Richie, Haixia Liu, Andrew Tseng, Wei Quan, Neal Ryan, David Brent, and Fuchiang R Tsui. 2022. Classifying social determinants of health from unstructured electronic health records using deep learning-based natural language processing. *Journal of biomedical informatics*, 127:103984.

Elham Hatef, Masoud Rouhizadeh, Iddrisu Tia, Elyse Lasser, Felicia Hill-Briggs, Jill Marsteller, Hadi Kharrazi, and 1 others. 2019. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR medical informatics*, 7(3):e13802.

Elham Hatef, Gurmehar Singh Deol, Masoud Rouhizadeh, Ashley Li, Katyusha Eibensteiner, Craig B Monsen, Roman Bratslaver, Margaret Senese, and Hadi Kharrazi. 2021. Measuring the value of a practical text mining approach to identify patients with housing issues in the free-text notes in electronic health record: findings of a retrospective cohort study. *Frontiers in public health*, 9:697501.

James Henderson and Mark Pearson. 2025. Privacy-preserving natural language processing for clinical notes.

Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéol. 2023. Can synthetic text help clinical named entity recognition? a study of electronic health records in french. In *The 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Brian Hutton, Ferran Catala-Lopez, and David Moher. 2016. The prisma statement extension for systematic reviews incorporating network meta-analysis: Prisma-nma. *Medicina Clínica (English Edition)*, 147(6):262–266.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Jitendra Jonnagaddala, Hong-Jie Dai, Pradeep Ray, and Siaw-Teng Liaw. 2015. A preliminary study on automatic identification of patient smoking status in unstructured electronic health records. In *Proceedings of BioNLP 15*, pages 147–151.

Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. 2020. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311.

Hamda Khan, Mathew Krull, Jane S Hankins, Winfred C Wang, and Jerlym S Porter. 2023. Sickle cell disease and social determinants of health: a scoping review. *Pediatric blood & cancer*, 70(2):e30089.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2023. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. *arXiv preprint arXiv:2307.12114*.

Amanda Lans, Laura N Kanbier, David N Bernstein, Olivier Q Groot, Paul T Ogink, Daniel G Tobert, Jorrit-Jan Verlaan, and Joseph H Schwab. 2023. Social determinants of health in prognostic machine learning models for orthopaedic outcomes: A systematic review. *Journal of Evaluation in Clinical Practice*, 29(2):292–299.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Chenyu Li, Danielle L Mowery, Xiaomeng Ma, Rui Yang, Ugurcan Vurgun, Sy Hwang, Hayoung Kim Donnelly, Harsh Bandhey, Zohaib Akhtar, Yalini Senathirajah, and 1 others. 2024. Realizing the potential of social determinants data: a scoping review of approaches for screening, linkage, extraction, analysis and interventions. *medRxiv*.

Jesse M Lingeman, Priscilla Wang, William Becker, and Hong Yu. 2018. Detecting opioid-related aberrant behavior using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1179.

Dmytro S Lituiev, Benjamin Lacar, Sang Pak, Peter L Abramowitsch, Emilia H De Marchis, and Thomas A

Peterson. 2023. Automatic extraction of social determinants of health from medical notes of chronic lower back pain patients. *Journal of the American Medical Informatics Association*, 30(8):1438–1447.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, and 1 others. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023a. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*, 30(8):1389–1397.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2021. Annotating social determinants of health using active learning, and characterizing determinants using neural event extraction. *Journal of Biomedical Informatics*, 113:103631.

Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. 2023b. The 2022 n2c2/uw shared task on extracting social determinants of health. *Journal of the American Medical Informatics Association*, 30(8):1367–1378.

Sanne Magnan. 2017. Social determinants of health 101 for health care: five plus five. *NAM perspectives*.

Ian Magnusson, Noah A Smith, and Jesse Dodge. 2023. Reproducibility in nlp: What have we learned from the checklist? *arXiv preprint arXiv:2306.09562*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

Montserrat Marimon, Aitor Gonzalez-Agirre, Ander Intxaurrondo, Heidy Rodriguez, Jose Lopez Martin, Marta Villegas, and Martin Krallinger. 2019. Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In *IberLEF@ SEPLN*, pages 618–638.

Elizabeth McNeill, Zoe Lindenfeld, Logina Mostafa, Dina Zein, Diana Silver, José Pagán, William B Weeks, Ann Aerts, Sarah Des Rosiers, Johannes Boch, and 1 others. 2023. Uses of social determinants of health data to address cardiovascular disease and health equity: a scoping review. *Journal of the American Heart Association*, 12(21):e030571.

Gaya Mehenni and Amal Zouaq. 2024. Ontology-constrained generation of domain-specific clinical summaries. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 382–398. Springer.

Danielle L Mowery, Brett South, Olga Patterson, Shu-Hong Zhu, and Mike Conway. 2017. Investigating the documentation of electronic cigarette use in the veteran affairs electronic health record: a pilot study. In *BioNLP 2017*, pages 282–286.

Hielke Muizelaar, Marcel Haas, Koert van Dortmont, Peter van der Putten, and Marco Spruit. 2024. Extracting patient lifestyle characteristics from dutch clinical text with bert models. *BMC medical informatics and decision making*, 24(1):151.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in english, french and spanish: masked language models outperform generative model prompting. *arXiv preprint arXiv:2402.12801*.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. Accessed: 2025-07-07.

J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema, and Paul E Stang. 2012. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60.

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, and 1 others. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12):2716–2727.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, and 1 others. 2023a. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Cheng Peng, Xi Yang, Zehao Yu, Jiang Bian, William R Hogan, and Yonghui Wu. 2023b. Clinical concept and relation extraction using prompt-based machine reading comprehension. *Journal of the American Medical Informatics Association*, 30(9):1486–1493.

Cheng Peng, Zehao Yu, Kaleb E Smith, Wei-Hsuan Lo-Ciganic, Jiang Bian, and Yonghui Wu. 2024. Improving generalizability of extracting social determinants of health using large language models through prompt-tuning. *arXiv preprint arXiv:2403.12374*.

Brian E Perron, Bryan G Victor, Gregory Bushman, Andrew Moore, Joseph P Ryan, Alex Jiahong Lu, and Emily K Piellusch. 2019. Detecting substance-related problems in narrative investigation summaries of child abuse and neglect using text mining and machine learning. *Child abuse & neglect*, 98:104180.

Suraj Rajendran and Umit Topaloglu. 2020. Extracting smoking status from electronic health records using nlp and deep learning. *AMIA Summits on Translational Science Proceedings*, 2020:507.

Swati Rajwal, Ziyuan Zhang, Yankai Chen, Hannah Rogers, Abeed Sarker, Yunyu Xiao, and 1 others. 2025. Applications of natural language processing and large language models for social determinants of health: Protocol for a systematic review. *JMIR Research Protocols*, 14(1):e66094.

Giridhar Kaushik Ramachandran, Yujuan Fu, Bin Han, Kevin Lybarger, Nicholas J Dobbins, Özlem Uzuner, and Meliha Yetisgen. 2023. Prompt-based extraction of social determinants of health using few-shot learning. *arXiv preprint arXiv:2306.07170*.

Shaina Raza, Elham Dolatabadi, Nancy Ondrusek, Laura Rosella, and Brian Schwartz. 2023. Discovering social determinants of health from case reports using natural language processing: algorithmic development and validation. *BMC Digital Health*, 1(1):35.

Russell Richie, Victor M Ruiz, Sifei Han, Lingyun Shi, and Fuchiang Tsui. 2023. Extracting social determinants of health events with transformer-based multitask, multilabel named entity recognition. *Journal of the American Medical Informatics Association*, 30(8):1379–1388.

Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, Farhad Nooralahzadeh, Lei Clifton, Laura Merson, David A Clifton, ISARIC Clinical Characterisation Group, and 1 others. 2024. Lightweight transformers for clinical natural language processing. *Natural language engineering*, 30(5):887–914.

Brian Romanowski, Asma Ben Abacha, and Yadan Fan. 2023. Extracting social determinants of health from clinical note text with classification and sequence-to-sequence approaches. *Journal of the American Medical Informatics Association*, 30(8):1448–1455.

Dmitry A Scherbakov, Nina C Hubig, Leslie A Lenert, Alexander V Alekseyenko, and Jihad S Obeid. 2025. Natural language processing and social determinants of health in mental health research: Ai-assisted scoping review. *JMIR Mental Health*, 12(1):e67192.

Fatemeh Shah-Mohammadi and Joseph Finkelstein. 2024. Utilizing rag and gpt-4 for extraction of substance use information from clinical notes. In *Collaboration across Disciplines for the Health of People, Animals and Ecosystems*, pages 94–98. IOS Press.

Pankaj Sharma, Imran Qureshi, and Minh Tran. 2022. Meta learning for few-shot medical text classification. *arXiv preprint arXiv:2212.01552*.

Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. 2018. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Rachel Stemerman, Jaime Arguello, Jane Brice, Ashok Krishnamurthy, Mary Houston, and Rebecca Kitzmiller. 2021. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. *JAMIA open*, 4(3):ooaa069.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, and 1 others. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Maxim Topaz, Ludmila Murga, Ofrit Bar-Bachar, Kenrick Cato, and Sarah Collins. 2019. Extracting alcohol and substance abuse status from clinical notes: The added value of nursing data. In *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 1056–1060. IOS Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

U.S. Department of Health & Human Services. 2003. Summary of the hipaa privacy rule. `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`. Accessed: 2025-07-22.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

Stella Verkijk and Piek Vossen. 2021. Medroberta. nl: a language model for dutch electronic health records. In *Computational Linguistics in the Netherlands*, volume 11, pages 141–159. Computational Linguistics in the Netherlands.

Yan Wang, Elizabeth S Chen, Serguei Pakhomov, Elliot Arsoniadis, Elizabeth W Carter, Elizabeth Lindemann, Indra Neil Sarkar, and Genevieve B Melton. 2015. Automated extraction of substance use information from clinical texts. In *AMIA Annual Symposium Proceedings*, volume 2015, page 2121.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. *arXiv preprint arXiv:2204.07705*.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, and 1 others. 2023. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *arXiv preprint arXiv:2311.09000*.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, and 1 others. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Elizabeth Geena Woo, Michael C Burkhart, Emily Alsentzer, and BK Beaulieu-Jones. 2024. Synthetic data distillation enables the extraction of clinical information at scale. medrxiv. *Published online September*, 28:2024–09.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. *arXiv preprint arXiv:2410.15702*.

Meliha Yetisgen and Lucy Vanderwende. 2017. Automatic identification of substance abuse from social history in clinical text. In *Artificial Intelligence in Medicine: 16th Conference on Artificial Intelligence in Medicine, AIME 2017, Vienna, Austria, June 21-24, 2017, Proceedings 16*, pages 171–181. Springer.

Zehao Yu, Cheng Peng, Xi Yang, Chong Dang, Prakash Adekkanattu, Braja Gopal Patra, Yifan Peng, Jyotishman Pathak, Debbie L Wilson, Ching-Yuan Chang, and 1 others. 2024. Identifying social determinants of health from clinical narratives: A study of performance, documentation ratio, and potential bias. *Journal of biomedical informatics*, 153:104642.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, and 1 others. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Yuan Zhao, Erica P Wood, Nicholas Mirin, Stephanie H Cook, and Rumi Chunara. 2021. Social determinants in machine learning cardiovascular disease prediction models: a systematic review. *American journal of preventive medicine*, 61(4):596–605.

Weipeng Zhou, Meliha Yetisgen, Majid Afshar, Yanjun Gao, Guergana Savova, and Timothy A Miller. 2024. Improving model transferability for clinical note section classification models using continued pretraining. *Journal of the American Medical Informatics Association*, 31(1):89–97.

Xinyu Zhou, Lovedeep Singh Dhingra, Arya Aminorroaya, Philip Adejumo, and Rohan Khera. 2025. A novel sentence transformer-based natural language processing approach for schema mapping of electronic health records to the omop common data model. In *AMIA Annual Symposium Proceedings*, volume 2024, page 1332.

A PRISMA-Based Paper Selection Methodology

A.1 Search Strategy

We conducted a systematic literature search inspired by the PRISMA framework. The search spanned February to June 2025 and included publications up to May 15, 2025. Figure 3 illustrates the article selection flowchart according to the PRISMA guidelines.

Databases searched and initial results. We constructed targeted Boolean queries to identify literature on SDOH extraction using NLP methods. For broad-scope databases such as Google Scholar, more restrictive queries were necessary to filter out unrelated results. In contrast, domain-specific repositories like the ACL Anthology, which primarily contain NLP-focused literature, required minimal filtering.

- **Google Scholar:** 1,590 results using `(SDOH OR “social determinants of health”) AND “extraction” AND (EHR OR “electronic health record”) AND (NLP OR “natural language processing”)`
- **PubMed:** 65 results using `SDOH AND NLP`
- **IEEE Xplore:** 412 results using `SDOH AND NLP`
- **ACL Anthology:** 505 results using `SDOH`
- **ACM Digital Library:** 23 results using `SDOH`

Total initial records: A total of 2,595 papers were gathered initially from different databases.

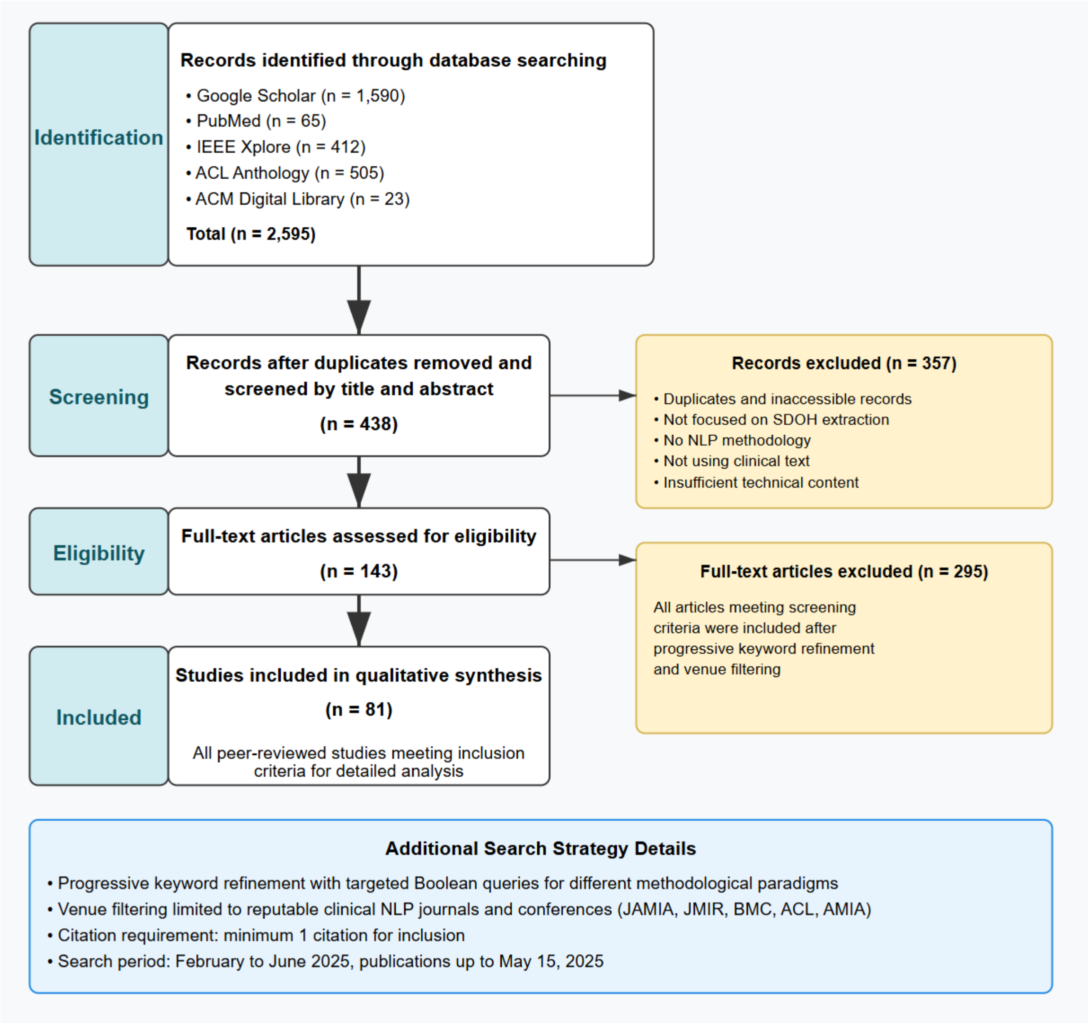

Figure 3: Article selection flowchart following PRISMA guidelines.

A.2 Screening and Eligibility Criteria

After removing duplicates and inaccessible records, we retained 438 papers for title and abstract screening. Removing duplicates, studies not focused specifically on SDOH extraction and NLP, or not using clinical text, and studies having insufficient technical content, 143 papers were selected for primary review. A total of 81 papers were finally selected after carefully going through the contents.

Venue filter: Only papers published in reputable venues related to clinical NLP with at least 1 citation were included. These included:

- **Journals**: JAMIA, JMIR, BMC, *npj* Digital Medicine.
- **Conferences**: ACL, AMIA.
- **Preprints**: arXiv, medRxiv, and other widely cited, clinically relevant preprint servers.

Inclusion criteria. We included papers that made methodological contributions to SDOH extraction using NLP techniques. Eligible studies focused on unstructured clinical text, such as free-text from electronic health records. Only those that offered empirical evaluation or meaningful technical insights were retained. Furthermore, we limited inclusion to papers that were either peer-reviewed or demonstrably cited in the research community.

Exclusion criteria. We excluded studies that relied solely on structured data or did not apply NLP-based methods. Opinion pieces, narrative reviews without new technical contributions, and papers that lacked citations or academic impact were also omitted. Additionally, we removed papers that were unrelated to clinical settings or did not explicitly address social determinants of health.

Category	Query (Google Scholar)	Results
Rule-based methods	(SDOH OR “social determinants of health”) AND extraction AND (EHR OR “electronic health record”) AND (NLP OR “natural language processing”) AND (“rule-based” OR “pattern matching” OR “dictionary-based” OR “regular expressions”)	665
Classical ML methods	... AND (“machine learning”)	1,870
Deep learning methods	... AND (“deep learning” OR “neural network”)	1,490
Transformer methods	... AND (“transformer” OR “pre-trained language model”)	504
LLMs	... AND (LLM OR “large language model”)	407
SDOH ethics-focused	... AND (Ethics AND Bias AND Privacy)	1,170
Clinical ethics-focused	(“clinical extraction” OR “medical extraction”) AND (NLP OR “natural language processing”) AND (Ethics OR Bias OR Privacy)	132

Table 2: Refined queries and result counts for methodological coverage.

A.3 Progressive Keyword Refinement

To ensure comprehensive methodological coverage, we issued targeted Boolean queries corresponding to distinct modeling paradigms for the **Google Scholar** database, including rule-based systems, classical machine learning, deep learning, transformer models, large language models, and ethics-related studies, as shown in Table 2. Each category was screened for relevance, technical rigor, and citation impact. After this multi-stage refinement process, we retained a final curated set of **81 peer-reviewed studies** for detailed analysis.

A.4 Data Extraction and Synthesis

From each selected paper, we extracted key metadata including title, authors, publication venue, and year. We also recorded task formulations and the specific modeling techniques used, along with dataset characteristics and their accessibility. Evaluation metrics and empirical results were cataloged to enable performance comparison. We additionally noted whether papers addressed ethical aspects such as bias, privacy, or hallucination. Lastly, we documented the limitations identified by each study and any proposed directions for future work. The extracted data were then synthesized thematically across modeling paradigms, dataset types, ethical considerations, and deployment practices to identify prevailing trends, challenges, and gaps in the literature.

B Dataset Information

Table 3 summarizes key details of existing SDOH datasets, including year of creation, data source, annotation granularity, covered SDOH categories, dataset size and type (e.g., notes, sentences), inter-annotator agreement (Cohen’s κ), and accessibility.

All datasets are access-controlled due to HIPAA regulations. Label distributions are skewed toward tobacco use, with factors like housing, childcare, and legal needs underrepresented. Inter-annotator agreement varies by dataset.

C Implementation and Training Configurations

C.1 Rule-based System Performance.

In the i2b2 2008 Smoking Challenge, rule-based systems achieved micro-F_1 scores ranging from **0.80 to 0.89** (Uzuner et al., 2008). Wang et al. reported F_1 of **0.89** for nicotine use detection using similar techniques but noted reduced scores for complex attribute extraction tasks.

C.2 Feature-engineered Classifier Performance.

Lingeman et al. trained a linear SVM with hand-crafted and sentiment features, achieving approximately **81% accuracy** for detecting opioid-related aberrant behavior from notes at the *University of Massachusetts Medical Center*. Despite this, such models still required expert-designed features and struggled to generalize across institutions or note

Dataset	Yr.	Source	Gran.	SDOH Categories	Size	IAA	Access/Limitations
i2b2 NLP Smoking Challenge (Uzuner et al., 2008)	2008	Partners HealthCare	Note	Smoking status (5)	502 notes	0.84 κ	Restricted
MIMIC-III (Gehrmann et al., 2018)	2018	MIMIC-III	Note	10 phenotypes (obesity, chronic pain, *etc.*)	1 610 notes	0.71–0.95 κ	Restricted (MIMIC)
CUMC Corpus (Feller et al., 2020)	2020	CUMC	Note (semi-sup.)	30+ factors (alcohol, housing, . . .)	4 663 notes	0.736 κ	Restricted
Wang15 (Wang et al., 2015)	2015	MTSamples, UPMC	Fine	Substance use (alcohol, drug, tobacco)	691 notes	0.80–0.93 κ	Restricted
(Yetisgen and Vanderwende, 2017)	2017	MTSamples	Fine	Substance abuse (7 dims.)	516 reports (1234 sents.)	0.59 F1 (initial)	Restricted
SHAC (Lybarger et al., 2021)	2021	MIMIC-III, UW	Fine	12 cats. (substance, employment, living, . . .)	4 480 sections	0.61–0.97 κ	Restricted (MIMIC)
(Han et al., 2022)	2022	MIMIC-III (SW)	Fine	13 cats. (SNOMED-CT / DSM-IV)	3504 sents.	0.70 agr.	Restricted (MIMIC)
(Lituiev et al., 2023)	2023	Low-back-pain notes	Fine	7 domains + mental health, pain	626 notes	0.95 agr.	Restricted
(Raza et al., 2023) Raza et al.	2023	LitCOVID API	Fine	10 cats. (gender, employment)	4000 case reports	0.75 κ	**Public** (LitCOVID) but Annotation Restricted

Table 3: Key SDOH datasets, their characteristics and accessibility.

structures, motivating the adoption of deep learning.

C.3 Bi-LSTM+CRF with Pre-trained Embeddings.

In the 2022 n2c2/UW shared task, BIOCLINICALBERT embeddings were integrated into Bi-LSTM+CRF pipelines to strengthen performance. These embeddings, derived from MIMIC-III discharge summaries and PubMed abstracts, provided contextualized representations tailored to the clinical domain. The resulting system reduced the performance gap to transformers to under five F_1 points while maintaining interpretability.

C.4 GatorTron Variants.

GATORTRON-MRC (345M parameters) framed SDOH extraction as a machine reading comprehension (MRC) task, using clinical prompts for question-style information retrieval. In contrast, GATORTRONGPT-20B adopted a decoder-only architecture with prompt tuning, allowing adaptation without full fine-tuning. Both models were trained on multi-billion-token clinical corpora.

C.5 T5-style Architectures.

T5-LARGE, used with constrained decoding, adopted a sequence-to-sequence format where the model generates structured output directly from the input transcript. This approach reduced post-processing and improved accuracy on SHAC, reaching $F_1 = 0.90$ (Romanowski et al., 2023).

C.6 PEFT Mechanisms.

Soft prompting or P-tuning modifies only the prompt embeddings while freezing the model backbone, minimizing the number of tunable parameters (Lester et al., 2021). *LoRA* (Low-Rank Adaptation) inserts rank-decomposed trainable matrices into transformer layers to adapt the model efficiently (Hu et al., 2022). These methods require orders-of-magnitude fewer resources than full-model fine-tuning, making them appealing for privacy-preserving clinical adaptation.

D Retrieval Augmented Generation (RAG)

Generative LLMs can hallucinate clinical facts, and they become expensive when a long note (often 10,000+ tokens) is fed in verbatim. **Retrieval-augmented generation (RAG)** tackles both problems. It first fetches the most relevant snippets from a vetted corpus (the patient's own record or an external Knowledge Base (KB)) and lets the LLM read only that compact context (Lewis et al., 2020).

D.1 Clinical performance

Cheetirala et al. (2025) and Jiang (2024) show that chunk level retrieval (top 4000 tokens) allows GPT

4o, Llama 2, and Mistral to match full note performance in surgical complication classification, with no significant drop in AUC, precision, recall, or F_1. The CLEAR pipeline (Lopez et al., 2025), a clinically guided retrieval method that filters notes using structured patient context and task specific cues, outperforms both embedding based retrieval and full note baselines. It achieves an F_1 of 0.90 while reducing input size from 6.1k to 1.1k tokens and inference time from 20.1 seconds to 4.9 seconds.

D.2 Synthetic supervision

RAG is still useless if you have *zero* labeled data. Woo et al. (2024) demonstrate that a 70B Llama-3 (Grattafiori et al., 2024) teacher can generate question–answer pairs that drive an 8B student to micro-$F_1 \geq 0.94$ on three clinical extraction tasks, all without exposing any protected text. Building on this approach for SDOH applications, Gong et al. (2025) demonstrate the value of synthetic data generation, using GPT turbo-0301 to create 2,280 synthetic clinical notes following domain expert annotation guidelines for SDOH categories, enabling robust evaluation across underrepresented social determinants like unstable housing. These approaches suggest a privacy-preserving path for SDOH extraction: large, cloud-only models can supply synthetic labels for social determinants; small, locally hosted models can then perform inference on real patient data.

E Bias, Fairness, Privacy, & Hallucination

We outline key ethical concerns for clinical LLMs (bias, privacy, and hallucination) along with real-world manifestations, potential harms, and common mitigation strategies in §3. Specific hallucination mitigation techniques are further described below.

E.1 Hallucination Mitigation Strategies

A number of techniques can be adopted to mitigate the risk of hallucinations by large language models.

(i) Grounding via retrieval. Conditioning the LLM on retrieved passages, RAG halved token usage with almost no loss in accuracy (Cheetirala et al., 2025). In another medical-QA scenario, RAG increased correct references from 20% to 55%, substantially reducing hallucinated evidence (Gilson et al., 2024).

(ii) Domain-specific tuning and calibrated decoding. Recent research in medical LLMs shows that domain-specific fine-tuning and calibrated decoding can significantly reduce hallucinations. For example, Xu et al. (2024) introduced *Alternate Contrastive Decoding* (ALCD) in medical information extraction tasks. ALCD applies contrastive decoding during inference, alternating between identification and classification objectives to suppress spurious token generations, substantially reducing factual errors compared to standard decoding methods. Similarly, Mehenni and Zouaq (2024) proposed an ontology-constrained decoding approach for clinical summarization. By integrating domain ontologies to guide the decoding process, the model restricts output to medically valid terms and relations, yielding more accurate and hallucination-free summaries on MIMIC-III.

(iii) Human-in-the-loop fact-checking. Wang et al. (2023) introduced Factcheck-GPT, a structured fact-checking pipeline following their own multi-stage Factcheck-Bench framework. The system decomposes LLM-generated text into atomic claims, retrieves supporting evidence, assesses stances, and produces verified responses. Factcheck-GPT achieved superior performance compared to existing tools like FacTool and FactScore on their benchmark, recording an F_1 score of 0.63 for claim-level verification and demonstrating broad improvements across sentence and document-level accuracy. Its fine-grained evaluation allowed tracing and correcting errors at each intermediate stage, significantly improving reliability over previous black-box approaches.

F Cross-Institutional Generalization Techniques

A few techniques have been adopted by recent clinical NLP extraction tasks. These techniques have proven to be effective and, therefore, can be adopted by future SDOH extraction works.

(i) Continued pre-training (CPT). Continued pretraining refers to further training a general biomedical language model on target domain text using masked language modeling. This includes *Domain Adaptive Pretraining (DAPT)*, which uses unlabeled text from the target domain, and *Task Adaptive Pretraining (TAPT)*, which uses unlabeled task-specific data. These approaches require no new annotations but significantly improve downstream performance by aligning the model more closely with the linguistic patterns of the deployment setting.

In cross-institutional SOAP section classification, Zhou et al. (2024) showed that applying both DAPT and TAPT raised the micro F_1 score from 0.756 to 0.808 across three datasets. Similarly, Gururangan et al. (2020) demonstrated that DAPT alone improved biomedical information extraction, increasing micro F_1 from 0.819 to 0.842 on the CHEMPROT dataset and from 0.872 to 0.876 on the RCT dataset. These results highlight the effectiveness of continued pretraining in modeling domain-specific language with minimal supervision.

(ii) Invariant-representation learning. Adversarial training can enforce invariant representations in biomedical NLP by exposing models to input perturbations such as typos, character swaps, and synonym substitutions. Araujo et al. (2020) show that fine-tuning BERT on both clean and adversarially modified data restores up to 20–23% performance lost to these perturbations, indicating improved robustness to superficial variations in medical text.

(iii) Lightweight meta-learning. Meta-learning techniques adapted for medical text classification can achieve data-efficient adaptation with limited examples. Sharma et al. (2022) report that their meta-learning model, combined with distributionally robust optimization, improves worst-case loss across disease codes and achieves performance comparable to few-shot language models when trained on medical note data. Meanwhile, Rohanian et al. (2024) present compact transformers (15–65M parameters) built via knowledge distillation and continual learning. These models perform on par with larger BIOBERT and CLINICALBIOBERT models and significantly outperform other small models on tasks such as named entity recognition (NER), relation extraction, inference, and sequence classification.

G Expanding to Multilingual SDOH Extraction

Technical advances in English SDOH extraction do not generalize globally. Nearly all benchmarks are English-only, even though over half of EHRs worldwide are not. This language imbalance risks reinforcing disparities in care delivery and data-driven tools.

(i) Data scarcity outside English. Non-English clinical corpora with SDOH labels are rare. MEDDOCAN (Spanish) (Marimon et al., 2019) and DRAGON (Dutch) (Bosma et al., 2025) include some de-identification and clinical annotations but lack social-history categories, making them unsuitable for supervised SDOH tasks. Researchers often resort to machine translation or cross-lingual transfer, which may compound biases.

(ii) Vocabulary and template mismatch. Many SDOH terms reflect US-specific institutions (e.g., "FOOD_STAMPS", "HOUSING"). These do not translate directly and often result in incoherent mappings in other languages. Label taxonomies built around English contexts fail under naive translation, requiring culturally aligned adaptation.

(iii) Limited language coverage in domain LLMs. Most domain-specific LLMs are English-based. Naguib et al. (2024) show that French and Spanish models trained on native data consistently outperform cross-lingual and zero-shot English models, indicating that language-specific modeling is necessary.

G.1 Approaches

A few solutions to these problems have been explored for other similar clinical NLP tasks. These are highlighted here:

(i) Translate-train → original-test. Fontaine et al. compare two cross-lingual approaches for clinical NER in French and German: (i) cross-lingual transfer using a multilingual model fine-tuned on English data, and (ii) translation-based methods, which either translate English training data into the target language ("translate-train") or translate target-language text into English ("translate-test") before extraction. They release a new French clinical NER test set (MedNERF) and show that both approaches achieve comparable F_1 scores $\sim (0.78 - 0.79)$, with careful design of translation pipelines.

(ii) Continued pretraining on local notes. MedRoBERTa.nl (Verkijk and Vossen, 2021), further pre-trained on Dutch clinical data, achieves strong macro-F_1 scores: 0.93 (smoking), 0.79 (alcohol), and 0.77 (drugs). These outperform ClinicalBERT translated to Dutch, which scored 0.92, 0.80, and 0.61, respectively (Muizelaar et al., 2024). Results, however, come from a single institution and need broader evaluation.

(iii) Multilingual synthetic data. Hiebel et al. (2023) trained NER models on French EHR cases generated using GPT-style clinical models. These models matched the performance of real-data–trained models, suggesting that synthetic data may help bootstrapping in low-resource settings, though cultural and linguistic alignment must be verified.